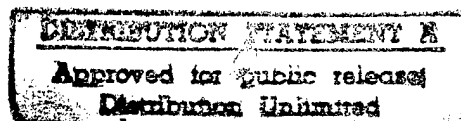A METHODOLOGY FOR THE ANALYSIS AND
PREDICTION OF AIR FORCE OFFICER
RETENTION RATES

THESIS

Mark A. Basalla, Captain, USAF

AFIT/GOR/ENC/96M-01

19961220 104

A METHODOLOGY FOR THE ANALYSIS AND PREDICTION

OF AIR FORCE OFFICER RETENTION RATES

THESIS

Presented to the Faculty of the School of Engineering

of the Air Force Institute of Technology

Air University

Requirement for the Degree of

Master of Science in Operations Research

Mark A. Basalla, B.S.
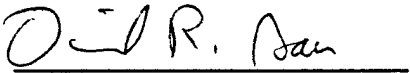
Captain, USAF

March 1996
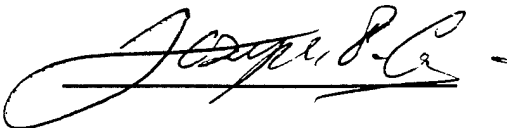
THESIS APPROVAL

STUDENT: Mark A. Basalla, Captain, USAF   **CLASS**: GOR-96M

THESIS TITLE: A Methodology for the Analysis and Prediction of Air Force Officer Retention Rates

DEFENSE DATE: 8 Feb 96

COMMITTEE	NAME/TITLE/DEPARTMENT	SIGNATURE
Advisor	David R. Barr Associate Professor Department of Mathematics and Statistics	
Reader	Joseph P. Cain Associate Professor Department of Operational Sciences	

Abstract

The purpose of this study is to investigate the effects of certain national economic conditions and certain Air Force related conditions on officer retention rates and to build, verify, and validate a multivariate linear regression model to be used by Air Force personnel management officials that will predict officer retention rates for rated and non-rated line officers aggregated by Yeargroups and AFSC groups. Previous retention models were reviewed to study possible predictors and methodologies.

Since retention can be seen as a binary (stay/leave) decision, the model of choice for a binary dependent variable is the multivariate logistic regression model. The logit transformation was used on this model for simplification. D.R. Cox gives three assumptions, that were valid in this case, so ordinary least squares can be used to estimate the parameters of the logit model.

The tournament approach of the Modified Miller's Method was used for variable selection. This new approach was first validated by computer simulation and then used in the model building process for all of the models in this effort. The output of this tournament approach was the model of choice for each AFSC and Yeargroup. The results of the individual AFSC models were not that good, so two-way without replication ANOVA was done in order to combine like AFSC's into several groups. There were six groups in all. A separate model was then built for each of the six groups. These results were much better.

Validation tests were performed with the fiscal year 1994 and 1995 data. In each test, the 90 percent prediction interval contained the actual retention rate for each AFSC group.

Table of Contents

List of Illustrations

A METHODOLOGY FOR THE ANALYSIS AND PREDICTION

OF AIR FORCE OFFICER RETENTION RATES

I. Introduction

Background

Since the beginning of the all volunteer Air Force in 1974, personnel officials have been interested in the answer to one key question: Why do some officers in the Air Force leave and others stay? During the Reagan Defense buildup there was an overabundance of officers experienced, however, so that this was a minor concern, until recently. If something in the economy changed, and many officers in a particular Air Force Specialty Code (AFSC) left, the managers would simply move people around from an overabundant AFSC to the AFSC that was short or bring in new officers to fill these vacancies. This was especially true for the non-rated line officer AFSC's.

The managers of the rated line officers, pilots and navigators, could not use the cross flow technique of filling voids because of the expensive training costs and long training times. These managers have always been interested in the relationship between economics and pilot/nav retention. This is evident in the many articles, research papers, and theses done on the cause and effects of pilot retention.

In the early 1990's, the whole Department of Defense, and the Air Force in particular, was in the middle of a congressionally mandated force drawdown. During this drawdown process, the Air Force implemented various methods to get certain numbers of officers in certain AFSC's to leave. Although each armed service was allowed a period of three to four years to effect their individual drawdowns, the Air Force chose to "front load" their losses in the first several years. These methods include voluntary separation incentives (VSI), selective separation bonus (SSB), temporary early retirement authority (TERA), and an involuntary reduction in force (RIF). The VSI/SSB are both voluntary separation programs in which money is the incentive to get the officer to leave. Both programs have targeted Yeargroups in specific AFSC's. The only difference is that the VSI has yearly payments of a certain amount for twice the number of years of service and the SSB has a lump sum payment based upon a mandated formula. TERA is the fifteen-year retirement program. Historically, officers can retire their commissions with full benefits after twenty to thirty-five years of service. The TERA makes it possible for certain officers to retire at the fifteen-year point with the same benefits, but with a minor deduction in retirement pension payments. RIFs are the most unpopular of all of the reduction programs because they are involuntary. The RIF in 1992 was where a board of senior officers, much like a promotion board, was held at Headquarters Air Force Military Personnel Center (HQ AFMPC). This board decided which officers in certain Yeargroup/ AFSC categories would stay and which ones would be forced to leave the Air Force.

Because of this drawdown, all of the AFSC managers are highly interested in the effects of certain national economic conditions and certain Air Force personnel policies on officer retention. They no longer have a large pool of officers to fill their voids. They also no longer have the large amount of money budgeted for training to cross train officers from one AFSC to another. Officer retention is a big issue to all of the AFSC managers because they still have to fill the positions required for day to day worldwide operations and they want to fill them with the best people they can.

It is very important to mention that officer turnover is a necessary and an intrinsically good thing. The type of desired turnover occurs when the less productive/skilled officers leave in a controlled and managed fashion. However, officer attrition can pose a dilemma when highly trained, highly skilled, and very useful officers leave in a uncontrolled and unpredictable manner thereby leaving mediocre officers to do the jobs.

Before the drawdown, the non-rated line officer retention rates were predicted merely by the simple average of past retention rates. However, since the drawdown, things have changed dramatically. Many officers live in fear of another drawdown with another potential RIF. The number of people who do not look to make the Air Force a career has increased dramatically. As more of the retirement benefits slowly decrease, this number could increase even more. Those that want to stay in the Air Force believe that they will not be allowed to stay in until retirement. What does this mean to the AFSC managers? Previous historical rates do not predict today's rates very well. This

failure of predictive capability is happening at a critical time when valid loss estimates are needed for good managerial decisions on future Air Force force structure.

The need for a better way to analyze and predict future officer retention rates is one of the vital issues addressed in this thesis research.

Problem Statement

The objectives of this thesis effort are: (1) investigate the effects of certain national economic conditions and certain Air Force related conditions on officer retention rates, (2) build, verify, and validate a multivariate linear regression model to be used by Air Force personnel management officials that will predict officer retention rates for rated and non-rated line officers aggregated by Yeargroups and AFSC groups.

Scope

All officers entering the Air Force receive at least a four-year active duty service commitment (ADSC). Historically speaking, most officers that have completed fifteen years will wait out the final five years and retire at the twenty-year (or later) point. For this reason, part of this research will concentrate on those non-rated line officers that have completed a minimum of four but not more than fifteen years of military service (YOS) in the USAF. The navigators (AFSC 12XX) incur a five year ADSC, so for them I will be concerned with 5 to 15 Yeargroups. The ADSC for the pilots (AFSC 11XX) has changed over the years. For 1986 and before, the ADSC was six years, for 1987 it was seven years, for 1988 to present it is eight years. Therefore, for the pilots, I will be

4

concerned with the 6 - 15 Yeargroups. The model will be built entirely on voluntary retention data. Therefore, if a particular Yeargroup in a particular year still is under the initial ADSC, then that data will not be used in the model building process.

This research effort will look at officers aggregated to the Yeargroup level with each Yeargroup split out by the first two digits of the AFSC. For example, we will lump al 11XX pilots for the 1987 yeargroup together. The reasoning for this aggregation is threefold. First, the data requirements needed at the individual entity level of detail for such a model would be enormous, secondly, the model would be too costly for use and high maintenance, and lastly, it is inappropriate.. Additionally, this researcher does not want to get involved in the politics of the relationship of gender, race, religion, and commissioning source with retention, should they exist.

II. Literature Review

Retention Models

Various approaches exist for modeling retention rates. One frequently used approach is to model the stay/leave decision of an individual by investigating, analyzing, and studying the impact of certain economic and personnel policies on this decision. A different approach is to aggregate various officers into specified categories and subsequently analyze the impact of these same economic and personnel policies on their aggregated retention rate. A retention rate is simply defined as the ratio of those who stay divided by the total number in that particular group for a particular period of time. For example, if 100 officers start a period of service and 83 stay through to completion four years later, we would calculate the retention rate as $83/100 = .83$. As previously mentioned, many earlier research efforts have concentrated on Air Force officer retention rates. Most of these studies dealt exclusively with pilot retention, but a few examined the total officer corps. This section of chapter 2 will detail some of these past research efforts.

The first model, built by Gotz and McCall of the RAND Corp., calculates a probabilistic weighted average of the cost of leaving the Air Force over various time horizons [Gotz and McCall, 1984:1-10]. This dynamic programming model looks at the individual officer's history, econometric data, and alternative personnel policies and wage compensations. It subsequently predicts a stay/retire decision at each time interval.

6

Their theory is that as long as this cost of leaving remains positive, then the officer would stay in. However, when the cost of leaving became zero or even negative, then the officer would leave. Maximum likelihood determined the values of the parameters of this model. A major downfall of this model is that it is focused at the individual level. It requires each of the officer's personal history (sex, age, etc.), entire past Air Force history (commissioning source, rank, years of service, etc.), and various predictions or historical data for the econometric variables for the time period that the stay/retire decision prediction is to be made. These requirements introduce a tremendously large data requirement of the entire officer corps.

The next model, built by Simpson, is a ordinary least squares (OLS) multivariate linear regression model which concentrates specifically on pilots in the seven to eleven years of service groups [Simpson, 1987: 12-21]. Simpson studied four different models (econometric, pay, profit, and job) and choose the best performing as his final model, the pay model. This model has three independent variables (civilian airline hires that year, corporate profits for the previous year, and the total unemployment rate for the previous year) that predict the logarithmic transform of the specific retention rate for the next year. The logarithmic transformation equation:

$$trate = -\ln(1 - rate + .001) \hspace{3cm} 1.$$

Trate is the new transformed dependent variable and rate is the original retention rate. It should be noted that the .001 is an entirely arbitrary number that is added to avoid the ln(0). A major flaw of this model is that the use of OLS regression implies that the dependent variable can take on any value between negative infinity to infinity. This is

truly not the case for this model. Since the rate is bounded between zero and one, then the transformed rate is bounded between -.001 and 6.91. A normal model can still be useful in the case where the interval [-.001, 6.91] contains almost all of the probability (e.g., we could use a normal model when $P[-.001 < trate < 6.91] > .99$). I did, however, use some of Simpson's independent variables in the initial stages of my model building process.

In 1990, Guzowski did follow on research to Simpson's thesis [Guzowski 1990:15]. In his research effort, Guzowski used Simpson's technique and variables to build a model that would predict the retention rates of pilots three years in advance. His effort, however, encountered problems similar to Simpson's effort.

A multivariate probit model was used by Roth, in 1981, to predict pilot retention rates [Roth 1981:87-107]. This effort concentrated on the decision of an individual pilot. He looked at the pilot's background, personnel file information, and military record for insights into the stay/leave decision making process. Again, the drawback to this model is that it concentrated on the individual and was therefore highly data intensive.

The Air Force Human Resources Laboratory published three studies on Air Force Enlisted Personnel retention rates that studied retention prediction in a new way. The authors of these are DeVany - May 1978, Saving - June 1980, and Saving - July 1985. These studies used accession and retention models of the enlisted force and a life cycle model of the enlistment force. The accession and retention models are stochastic process models in which the force requirements were viewed as the number of servers and the mean time in the Air Force was viewed as the service time. The results of these two

models were then used as input into the life cycle model. The theory of the life cycle model was that the enlisted individual will make decisions based on the present value of future earnings from either the military or civilian sector. Although these studies look at retention in a very interesting way, they also focus on the individual level.

The final model, built by Cromer and Julicher, is an OLS multivariate linear regression model [Cromer and Julicher 1982: 28-37]. They studied the effects of several econometric variables on pilot retention rates. Three models built included a no lag, a 6-month lag, and a 1 year lag of econometric variables. They also looked at using factor analysis on the econometric variables to capture all of the relevant information while discarding the irrelevant information. Their model of choice was the one with the data unlagged. The negative aspect of this model is that the pilots are not split out by years of service. This model, therefore, assumes that the pilot retention rates are the same across the years of service spectrum, which is clearly a sweeping generalization. An additional problem with this model is that the bound on the dependent variable is between zero and one.

The models described above provided insight for a methodology and guidance in the choice of independent predictor variables to be identified, analyzed, and discussed in this thesis effort. Although most of these models were based on the individual pilot's decision making process, the methodology can be expanded to the group level for all of the Air Force officers. The methodology of this thesis effort will be discussed in chapter three of this document.

Modified Miller's Method

The models built by Simpson, Guzowski, and Julicher/Cromer used the stepwise procedure of OLS regression model building. The stepwise procedure is an iterative search method that develops a sequence of regression models, where at each step an independent variable is either added or deleted. The criterion for adding or deleting a variable is in terms of the error sums of squares reduction [Neter and others, 1990:453]. This method, along with most other model building methods, has a potential problem of overfitting the data. Alan J. Miller, the Senior Principal Research Scientist of the Commonwealth Scientific and Industrial Research Organization (CSIRO) Division of Mathematics and Statistics in Melbourne Australia, states:

> In general, it is gross optimism to hope that an *ad hoc* procedure of adding one variable at a time, and perhaps plotting residuals against everything which comes to mind, will find the best fitting subset [Miller, 1990:13].

In 1984, Miller described a method to overcome this problem of data overfitting by augmenting the variable pool with a number of extraneous random variables and using them to determine which variables originally in the model might also be extraneous [Miller, 1984,1990:84]. In 1993, Woollard applied this procedure and named it Miller's Method [Woollard, 1993:21-22]. Woollard augmented the variable pool with an equal number of extraneous random variables, generated from the standard normal distribution, and the forward selection procedure was then used. The stopping criteria for this procedure is the inclusion of the first known extraneous random variable. The preferred model was the model built previous to the inclusion of the first known extraneous random variable. The variables not included were considered insignificant.

10

Mutlu observed in 1994 that in reality, Miller's Method was a simulation of extraneous variables, because randomly generated known random variables are exposed to the same sampling error that the extraneous variables are exposed to [Mutlu, 1994:41-42]. Like other simulation techniques, Miller's Method should be run multiple times, in order to minimize the effect of this sampling error. Mutlu named this technique of multiple runs the Modified Miller's Method (MM Method) [Mutlu, 1994:41]. In his research effort, Mutlu had to answer three fundamental questions for the use of the MM Method.

1.) How many known extraneous variables should be augmented to the variable pool?
2.) How many simulation runs of MM Method should be run?
3.) How do you choose the best model?

Mutlu's answers to these questions follow [Mutlu, 1994: 43-57].The answer to the first question is dependent on the sample size of the data. The number of known extraneous random variables is equal to the augmentation coefficient multiplied by the number of possible predictors. For a sample size of 10,20, 30 and larger, the augmentation coefficient is 1, 1.5, and 2 respectively. The number of simulation runs is given by this formula:

$$Runs > 9\left(\frac{V}{R}\right) \qquad\qquad 2.$$

where V = Total number of variables,

R = Number of known extraneous random variables.

The procedures for the selection of the most appropriate model is to first find the most frequent model. If it is not the null model (that is, the model without a single variable),

11

then this one is the best model. If it is the null model, then calculate the theoretical

probability of getting the null model assuming all of the variables are extraneous.

$$P\ (Null) = \frac{R}{V} \qquad 3.$$

with V and R as defined above. If the calculated frequency of the null model is much

less then the theoretical probability, then choose the second most frequent model. If the

calculated frequency of the null model is much more then the theoretical one, then the

null model is chosen. If it is close, then look at how many deviations the calculated value

is from the expected value, where

$$E\ (Null) = Runs \ast P\ (Null) \qquad 4.$$

and
$$V\ (Null) = Runs \ast P\ (Null) \ast (1 - P\ (Null)) \qquad 5.$$

Hence, good judgement is necessary in determining whether the null model is the best

model.

Logit Models

A stay/leave decision could be considered as a dichotomous outcome, also called

a quantal or binary outcome. The logistic distribution is commonly used when an

outcome is binary. The form of the multivariate logistic distribution is:

$$\pi\ (x) = E\ [y \mid x] = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}} \qquad 6.$$

where $0 \le \pi\ (x) \le 1$ and x is the vector $[x_1 \ldots x_p]^T$. A very important transformation of

$\pi\ (x)$ is the logit transformation. The logit transformation is defined as:

$$logit = g\ (x) = \ln \left(\frac{\pi\ (x)}{1 - \pi\ (x)} \right) \qquad 7.$$

12

which simplifies to,

$$logit = g(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad\qquad 8.$$

The importance of the logit is that it has many of the desirable properties of a linear

regression model. The logit is linear in its parameters, may be continuous, and may

range from negative infinity to infinity, depending on the range of x [Hosmer and

Lemeshow, 1989:7]. In fact, Cox in 1970 states that the method of ordinary least squares

can be used to give asymptotically efficient procedures for the logit transformation when

binary observations are grouped into sets where all trials in the same set have the same

probability of success [Cox, 1970:30-32]. Each set is assumed to contain a reasonably

large number of equal, or nearly equal, trials. In the case of officer retention, the first

assumption implies that each officer in the group under the same influences will have the

same probability to stay in the Air Force. For this research effort, this will be assumed to

be valid. The second assumption of a large number of equal, or nearly equal trials

implies that the sample size in each group is relatively large and constant. The large

sample size will be discussed in the data collection section of the next chapter. The equal

number of trials requires good analytical judgement. If the trials are not nearly equal, the

model will have heteroscedasticity. The remedy for this would be the use of weighted

least squares. Therefore, for this effort, the trials will be assumed to be nearly equal, and

OLS regression will be used for the model building process. The residuals were checked

for heteroscedasticity. If heteroscedasticity existed, the model building process would

have been redone using weighted least squares regression.

It is evident that equation 7 needs a slight modification because it will be
undefined when $\pi(x) = 0$ or 1. Cox defines a modified logit transformation as [Cox,
1970: 33]:

$$logit = g(x) = \ln\left(\frac{\pi(x) + .5}{1 - \pi(x) + .5}\right) \qquad 9.$$

Note that the .5 plays a role similar to that of the .001 in equation 1 (p. 7). For the
remainder of this research effort, $\pi(x)$ will be the retention rate of the specific group
under study.

III. Methodology

Introduction

The purpose of this chapter is describe the research procedures for this thesis effort. There are two parts to this chapter. The first deals with the task of empirically validating the tournament approach of the MM method of variable selection and the second discusses the procedure for building a model that attempts to predict Air Force officer retention rates. The model building procedure is split up into the following steps: 1) data collection, 2) independent variable screening and initial model building, 3) diagnostic residual analysis, 4) model revision, and 5) model verification and validation. This chapter will explain each of these steps.

Tournament Approach

One approach to using the MM method of variable selection is to place all of the possible predictor variables into the variable pool along with all of the known extraneous random variables. The MM method, as described in chapter 2 of this document, is then used on this large variable pool. There is no problem with this approach when the number of variables in the original model is small. When the number of variables in the original model is large, as it is here because of interactions, the computer resource requirements required becomes a significant problem. The essential variables for this research effort include 10 econometric, 1 military specific, and five sets of binary

variables (18 AFSC's, 11 YOS, RIF that year, VSI/SSB that year, TERA that year). A total of 6720 predictor variable combinations are possible (including interactions with the binary variables). The SAS system is not capable of handling that many variables with only 19 years of annual data.

A tournament approach to the MM method of variable selection is a method in which the total variable pool of the possible predictor variables and known extraneous random variables are partitioned into several mutually exclusive and collectively exhaustive subsets. It is important to note that the ratio of known extraneous random variables to possible predictor variables must remain constant throughout this variable selection process. The MM method is then employed on each of these subsets. The "winners" at this level are then aggregated into a few variable pools and the MM method is subsequently used again. This aggregation and use of the MM method is utilized until all of the winner variables can be combined into a single variable pool and then the MM method is used a final time with the winner variables being included in the model. This iterative process is very useful for large variable pools such as the size of the variable pool in this research effort.

Intuitively speaking, this tournament approach makes a lot of sense. If a variable is truly extraneous, then this approach should not include the variable in the final model. However, what was not so clear is what happens when collinearity between variables exists. In the stepwise regression procedure, the variable that has the highest predictive power will be included first, then it will only include variables that add a specified amount of additional predictive power to the model. A variable that is moderately

16

collinear with the first variable and which adds more than the certain amount of additional predictive power, then it will be included in the model. Usually, the more a variable is collinear with the first variable, the less chance it will have of being included in the model.

The task of empirically validating the tournament approach was accomplished by simulation. First a pool of 24 variables each with a sample size of 25 was built, 12 variables were valid predictors of the response and 12 were known extraneous random variables. One of the seven valid predictors was highly correlated (scalar multiples) with the remaining five non-valid predictors and the remaining six valid predictor variables were not correlated. The extraneous random variables are generated from a normal distribution with a mean of zero and a variance of 1 (N(0,1)). The model of the dependant variable is of the form,

$$y = 10x_1 + 5x_2 + 4x_3 + 3x_4 + 2x_5 + x_6 + .5x_7 + \epsilon \qquad \qquad 10.$$

where x_1: first of the highly correlated variables,

x_2 -x_7: the six uncorrelated variables,

ϵ: error $\sim$ N (0,1).

The 25 values of the dependant variable were computed using equation 10.

Next, the pool of 24 variables was split into four subsets with six variables each (three known extraneous and three possible). Each of these four subsets were processed through the MM method 18 times. The winner variables of each subset were then aggregated into one variable pool, and the MM method was ran an additional 18 times on this pool. The winner variables of this last run were the exact same variables that won

when I ran all 24 variables together through the MM method eighteen times. I then executed this routine six different times with different combinations of the twelve possible predictor variables. The outcome was always the same as when the whole variable pool was run. With this result, I concluded that the tournament approach to the MM method was indeed a valid technique. This approach is implied when the MM method is used for the rest of the research effort.

Data Collection

The first step of the model building process is data collection. A list of the econometric and military specific variables with each source is included in Table 3-1A & Table 3-1B.

Table 3-1A: Econometric variables and sources

Econometric Variables	Source
Airline Hiring Rates	FAPA
Leading, Coincident, and Lagging Indexes	BEA - Survey of Current Business
Consumer Sentiment Index	BEA - Survey of Current Business
Index of Help-wanted Advertising	BEA - Survey of Current Business
Federal Reserve Discount Rate	BEA - Survey of Current Business
Civilian Unemployment Rate	BEA - Survey of Current Business
Pay Difference (Employment Cost Index - Military Pay Increase Rate)	Monthly Labor Review & Armed Services Almanac
White Collar Unemployment Rate	Employment & Earnings

Table 3-1B: Military specific variables and sources

Military Specific Variables	Source
Advanced Academic Degree Percentage	AFMPC
Pilot Bonus Eligibility	AFMPC
RIF, VSI/SSB, and TERA Years	AFMPC

It should be noted that the econometric variables are published either monthly or quarterly. A simple average of the months/quarters in a particular Fiscal Year was calculated, and this average was used in the model building process. The fiscal year was used instead of the calendar year because all of the military specific data is in Fiscal years. These variables are used in the initial model screening phase of the building process. Specific definitions of each variable and reasons for possible inclusion in the model is found at Appendix A of this document. During the screening process, each of the variable's potential predictive power were analyzed and the variables that were chosen to be statistically related to the retention rate were included in the final model. Since most of this data is econometric, one would expect a severe problem of multicollinearity. However, one should recall that the stepwise procedure handles this problem by adding variables into the model one at a time. At each step of the process, the new variable is analyzed for statistical correlation to the dependent variable with all of the other included variables already in the model. The variance inflation factors will be analyzed as a safety measure for multicollinearity. For the remainder of this research effort, the data collected is assumed to be accurate.

It was decided that only AFSC's that had an average yeargroup size of 40 individuals or more would be considered for analysis at the individual AFSC level. This decision is to fulfill Cox's assumption of large sample size in each group. The other AFSC's will be grouped with other larger AFSC's. Table 3-2A contains the individual AFSC's that will be grouped later in the research effort. Table 3-2B contains the individual AFSCs that are considered, their average size, and a brief description.

Table 3-2A: AFSC's to be grouped later

AFSC	Description
10	Operations Group Commander
16	Air Attaché
20	Logistics Commander
22	Space and Missile Maintenance
30	Chief, Mission Support
34	MWR and Services
35	Public Affairs
36	Personnel Programs Officer
38	Manpower Management

Table 3-2B: AFSCs Considered

AFSC (1st Two Digits)	Average Size	Description
11	1190	Pilots
12	529	Navigators
13	321	Space, Missile, ATC
14	132	Intelligence Officers
15	62	Weather Officers
21	153	Maintenance Officers
23	61	Supply Officers
24	43	Transportation Officers
25	43	Logistics Officers
31	47	Security Police
32	90	Civil Engineers
33	294	Communication-Computer Systems
37	107	Information Management
61	60	Scientists
62	225	Engineers
63	80	Acquisition Officers
64	60	Contracting Officers
65	73	Financial Management

Variable Screening and Initial Model Building

Initially, three separate ANCOVA models were built for each AFSC. Each model contained each of the ten econometric variables , one military specific variable, ten indicator variables for the 11 YOS, and two more indicator variables if a RIF or

21

VSI/SSB/TERA occurred that year. AFSC 11 (pilots) also had one more indicator variable for pilot bonus eligibility. The only difference in the three separate models was that the ten econometric variables were lagged one, two, or three years respectively. Each of the models contained the cross products of the 13 indicator variables with the 13 quantitative variables and the cross products of 13 quantitative variables, the ten YOS indicator variables, and three exit bonus indicator variables. These variable pools were augmented with 240 known extraneous random variables that were drawn from the standard normal distribution.

The tournament approach of the MM method was used for variable screening on each of the three models for each AFSC. Eighteen runs of the forward stepwise procedure in SAS was considered valid for each iteration of the tournament. The winners from the three models were aggregated into one variable pool. This pool was augmented with an equal amount of fresh extraneous random variables. Eighteen runs of the forward stepwise procedure was used with this pool of merged variables. The winners from this merged variable pool run were included in the final model for each AFSC. Appendix B of this document contains the results of this variable screening and model merging for each of the separate AFSC's.

The results of the merged models were not good. In speaking with the personnel analysis division at AFMPC, it was decided that a "good" model would have a r-squared value between 0.4 and 0.7. Anything below this range would be marked as not good enough and anything above might be considered as unrealistic and overfitting the data. Six of the eighteen AFSC's had R-squared values above the .4 threshold. When

comparing the average size of the AFSC and the R-squared value, a visible pattern was readily evident. As the average size of the AFSC grew, so did the R-squared value. This pattern is the exact reason why the models were being built on the group level. The breakdown of the AFSC's by the first two digit's was making some of the groups too small. In essence, we were back to the prediction on the individual level.

A simple remedy to this problem is to aggregate the AFSC's into groups of like AFSC's. The ANOVA procedure with the Scheffe method for testing equality of means was used to determine which AFSC's were statistically alike. SAS was used for this procedure. Table 3-3 contains the groups of the AFSC's that could be combined both statistically and logically.

Table 3-3: AFSC Grouping

Group	AFSC
1	11
2	13,14,15
3	21,23,24,25,31
4	33,37,64,65
5	32,61,62,63
6	12

Each of the AFSC's in Table 3-2A (10,16,20,22,30,34,35,36,38) that were deemed too small to be analyzed by themselves were put into one of the six aforementioned groups. Since the average size of these AFSC's was too small for any

statistical significance testing, this dispersing was purely based on logic and the specific composition of these AFSC's. For example, AFSC 10 and 16 are primarily pilots, so they were put in group 1. Table 3-4 outlines the final AFSC groups.

Table 3-4: AFSC Grouping (all AFSC's)

Group	AFSC's
1	10,11,16
2	13,14,15
3	20,21,22,23,24,25,31
4	30,33,34,35,36,37,38,64,65
5	32,61,62,63
6	12

Diagnostic Residual Analysis

Diagnostic checking for model aptness is usually done through the examination of the residuals (e_i), where e_i is simply the difference between the observed value and the fitted value. It should be noted that there is a distinction between a residual and the true model error (ϵ_i). The true model error is the difference between the observed value and the expected value of the observed value and e_i is an estimate of ϵ_i. For an OLS regression model, the true model errors, ϵ_i, are assumed to be independent normal random variables, with a mean equal to 0 and constant variance of σ^2. If the fitted model is indeed apt for the data at hand, the observed residuals should then reflect the properties assumed for the true model error (ϵ_i) [Neter and others, 1990: 115]. There are three

24

areas of residual diagnostics that were analyzed. These include heteroscedastic (do not have a constant variance) residuals , residuals that are not independent, and not normally distributed residuals.

Heteroscedasticity is easy to detect by looking at a graph of the residuals versus the predicted values. Heteroscedasticity is present when the spread of the residuals is not equal across the graph, that is, as the predicted value increases, the spread either increases or decreases.

The Durbin-Watson statistic was used to indicate whether there is any pattern in the residuals. The D-W statistic is:

$$D - W = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}, \qquad 11.$$

where n is the sample size. If the residuals are essentially random (independent), then the D-W statistic is around 2. If there is positive autocorrelation in the residuals, the D-W statistic will be less then 2. If there is negative autocorrelation in the residuals, the D-W statistic will be greater than 2. The range of the D-W statistic is from 0 to 4, and the theoretical underpinnings of this statistic are complex, so that in practice, reference is made to tables for approximate significance tests [Makridakis, 1983: 53 - 54]. SAS will output the D-W statistic and the appropriate significance level.

There are two ways to check for the normality of the residuals. The first is to plot a normal probability plot of the residuals. In this plot, each residual is plotted against its expected value when the distribution is normal. A plot that is nearly linear suggests

agreement with normality, whereas a plot that departs substantially from linearity

suggests that the error distribution is not normal [Neter and others, 1990:125]. The

second way is to compute the Kolmogorov statistic and probability of getting a larger

statistic when the sample is in fact normal. The univariate procedure in SAS will conduct

both of these methods.

Model Revision

There are many ways to attempt to revise a OLS regression model if it shows

signs of either heteroscedasticity, autocorrelation, or non-normal residuals.

Heteroscedasticity can be addressed through the use of weighted least squares regression

or a transformation of the dependent variable. Autocorrelation can be addressed by

adding one or more independent variables, transforming one or more of the existing

variables, or even fitting a Box-Jenkins (ARIMA) model to the residuals. Non-normal

residuals are also usually addressed by transforming one or more variables. I will not go

into any further details on these procedures since the developed models' residuals did not

show any significant signs of heteroscedasticity, autocorrelation, or non-normality.

Model Verification and Validation

Model verification is a process to determine if the developed model predicts the

logit of a retention rate. This test is accomplished by simply getting data for any given

year and using it to make a prediction. If the inverse of the predicted logit is between

zero and one, then the model is verified at that point. Because of the method of

calculation, this inverse must be between zero and one. Sensitivity analysis can be

accomplished to determine if the model is verified across the entire range of each of the independent variables in the model.

Model validation is the process to determine if the fitted model accurately predicts the logit and thereby the retention rate of the given AFSC group in the given YOS. The models for each of the AFSC groups were developed using data from 1976 to 1993. The data from 1994 and 1995 was withheld from the building process specifically for the validation process. This data was then input into the models and prediction point estimates and the prediction intervals of the particular logit values were computed. These estimates and intervals were then transformed into the retention rate space. The researcher then checked if the prediction intervals contained the actual retention rate. The results and analysis of this interval check is in the next chapter.

IV. Findings and Analysis

Introduction

In this chapter, I shall discuss and analyze the selected model, the results of the diagnostic residual analysis, and the results of the model verification and validation process. For brevity, only the group 1- Pilots (AFSC 10, 11, 16) model will be discussed in this chapter. A brief discussion of the remaining AFSC groups is contained in Appendix C.

Selected Model for Group 1

The ANOVA table for the prediction model selected for Group 1 based on data from 1976 - 1993 is as follows:

Dependent Variable : Logit transformed retention rate

Source	DF	Sum of Squares	Mean Squares	F Value	Prob>F
Model	9	52.54090	5.83788	66.183	0.0001
Error	159	14.02504	0.08821		
C Total	168	66.56594			

R-square	0.7893	Adj R-sq	0.7774

Parameter Estimates

Independent Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
Intercept	1	5.239586	0.74559024	7.027	0.0001
AADper	1	2.268479	0.12926141	17.550	0.0001
Lead1	1	-0.042235	0.00738376	-5.720	0.0001

28

Urate1	1	0.107702	0.02806401	3.838	0.0002
Help1Y6	1	-0.008882	0.00138361	-6.419	0.0001
Plt1Y6	1	0.000309	0.00003873	7.976	0.0001
White1Y10B	1	0.208614	0.05680248	3.673	0.0003
Paydif1Y6B	1	0.476408	0.09066869	5.254	0.0001
Sent3	1	-0.009471	0.00360552	-2.627	0.0095
Plt3Y6	1	0.000138	0.00002989	4.626	0.0001

where   AADper :   the percentage of the pilots in that YOS that have a advanced academic degree,

        Lead1 :   Leading economic indicator for the previous year,

        Urate1 :   Unemployment rate for the previous year,

        Help1Y6 :   Cross product of the Help-wanted in advertisement index for the previous year and the YOS 6 indicator variable,

        Plt1Y6 :   Cross product of the Airline hiring rate for the previous year and the YOS 6 indicator variable,

        White1Y10B: Triple cross product of the white collar unemployment rate for the previous year, YOS 10 indicator variable, and the pilot bonus indicator variable,

        Paydif1Y6B: Triple cross product of the pay difference for the previous year, YOS 6 indicator variable, and the pilot bonus indicator variable,

        Sent3 :   Consumer sentiment rate from three years previous,

        Plt3Y6 :   Cross product of the Airline hiring rate from three years previous and the YOS 6 indicator variable.

As seen from the ANOVA table, all of the independent variables are significant. The model is also significant. The R-square value of .7893 is the best for all of the models. It should be noted that YOS 6 should only be used for years 1992 and before, YOS 7 for 1993 and before, YOS 8 & 15 can be used anytime. These restrictions on the YOS for pilots is due to the changing of the ADSC incurred for pilot training.

In order to make this model useful, the analyst must use the inverse of the logit transform in order to get the predicted retention rate. The inverse logit transform is:

$$Predicted\ Retention\ Rate = \frac{e^{Predicted\ logit}}{1 + e^{Predicted\ logit}} \tag{12.}$$

There are a few interesting things to note about the independent variables. First of all, the coefficients for the leading economic indicator and for the help-wanted advertising index are negative. This makes perfect sense since as either the leading indicator or the help-wanted index gets bigger, the better the economy is and the more pilots would feel safe in leaving the Air Force. The coefficients for the unemployment rate and the white collar unemployment rate are both positive. This again makes sense since higher rates for both of these would indicate a poorer economy and pilots would therefore feel safer in staying in the Air Force. The coefficient for the pay difference is also positive. This positive coefficient is very misleading. The definition of pay difference, military pay increase minus the employment cost index, is the answer to this confusion. If the ECI is larger than the military pay increase, then the pay difference is negative, thereby having a negative effect on retention. This makes sense again, since if the civilian sector is getting larger pay raises than the military sector, more pilots will want to leave, thereby making the retention rate fall. The positive coefficients of the two airline pilot hire variables seem counter-intuitive. One would expect these to have a negative impact on retention. However, since both of these are interaction terms (with YOS 6), this could be interpreted as lower rates for YOS greater than six (rates lower than YOS 6).

Diagnostic Residual Analysis

As mentioned in chapter 3 of this document, the way to check for heteroscedasticity is through the examination of the graph of studentized residuals versus the predicted value. As seen in Figure 4-1, it appears that the studentized residuals are

30

equally spread across the graph. Therefore, the assumption of homoscedasticity of the residuals is valid.

Figure 4-1 : Plot of Studentized Residuals vs. Predicted Value of Logit

```
      -+---------+---------+---------+---------+---------+---------+-
       |                                                           |
    3 +                                                            +
       |                                *                          |
       |                  *                                        |
       |                        *         *                        |
    2 +            *       *         *   *       *                 +
       |              *                *                           |
       |              *                        *                   |
       |              *           ** *             *               |
  S    |              *                                            |
  t 1 +          * *      ***   *                  *               +
  u    |            * *      ** ** *****                           |
  d    |        *  *  ***  **   **  **                             |
  e    |      * *  * * ***  * *   *   *                            |
  n    |                    * ** **                                |
  t    |                                                           |
  i 0 +----------------*-----*-**-*-----------------------------   |
  z    |          ** **    ** * **   ****    * *                   |
  e    |        *  *  *       * *  * *  *                          |
  d    |           *           *             * *                  |
       |              **            ***         *                  |
  R    |        *   *  *   **  *                                   |
  e -1 +      *  *  *     *       * **    * * *   * *               +
  s    |           *             *           * *                   |
  i    |         *               *                        *        |
  d    |                         * *                               |
  u    |                                                           |
  a    |         *  *  *    *                *                     |
  l -2 +           *                                               +
       |                                                           |
       |                  *           *                            |
       |                                                           |
   -3 +                         *                                  +
       |                                                           |
       |                                                           |
       |                                                           |
   -4 +                                                            +
      -+---------+---------+---------+---------+---------+---------+-
       0.5      1.0       1.5       2.0       2.5       3.0       3.5       4.0
                        Predicted Value of LOGIT
```
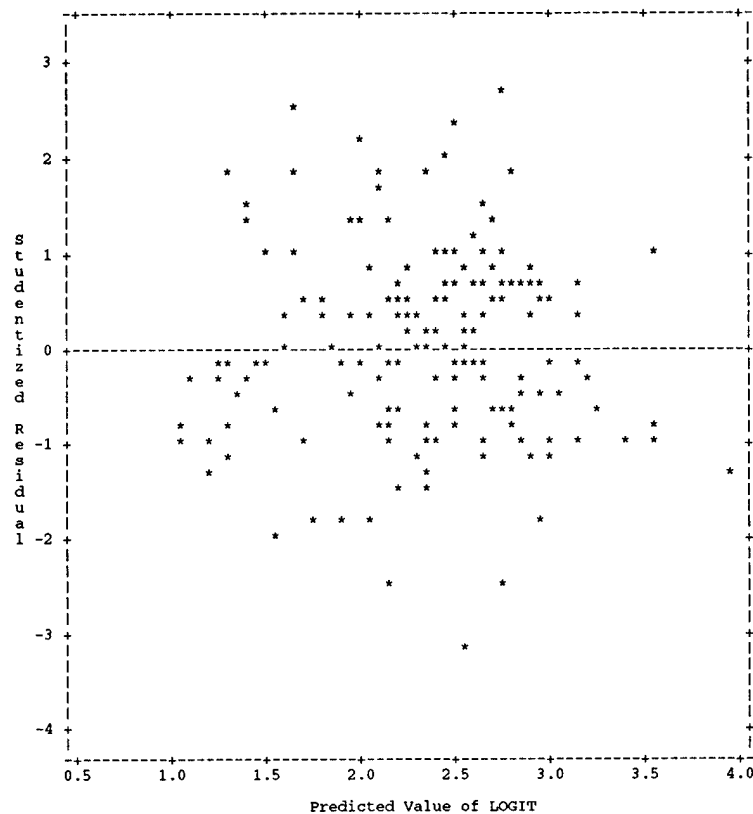
Next, the Durbin-Watson statistic was calculated in order to indicate if a pattern was present in the residuals. The calculated statistic was 1.995. Since this is very close to 2, it indicates that there is no significant pattern in the residuals and the assumption of independence is valid.

The final residual check is for normality. This was accomplished by first calculating the Wilkes-Shapiro statistic. The calculated statistic was 0.9889, with a 0.9070 probability of obtaining a value less then this when sampling from a normal

31

population. This indicates that the residuals are normal. The second check was by examining the normal probability plot of the residuals. As seen in Figure 4-2, it appears that the plot is nearly linear. This again suggest that the residuals are normal. Therefore, the assumption of normality is valid.

Figure 4-2 : Normal Probability Plot of the Studentized Residuals

```
2.75+                                              *
    |                                         **+**+
    |                                       ****+
    |                                    *****
    |                                  ******
    |                               ******
-0.25+                           ******
    |                          ******
    |                      ******
    |                   *****+
    |  ++*++
    |+  *
-3.25+*
    +----+----+----+----+----+----+----+----+----+
       -2        -1        0        +1        +2
```
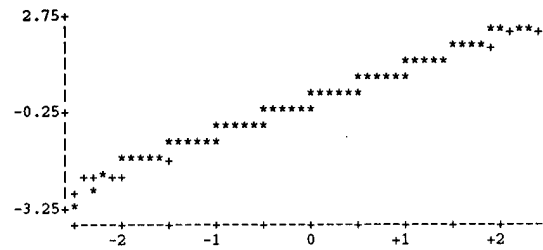
<u>Model Verification and Validation Results</u>

The entire set of data points for years 1975-1993 were imputed into the model shown above. The predicted retention rates were calculated using the inverse logit transformation above. All of these predicted rates fell between zero and one. This is no surprise since the mathematical derivation of the logit transformation insures that the inverse logit will fall between zero and one. The model is therefore verified for use in predicting retention rates given the needed independent variable inputs.

The FY 1994 data that was held back was imputed into the model described above using SAS. The output was logit point estimates and the lower and upper bounds for a 95% prediction interval for each of the YOS. Each of these corresponding values was then transformed into retention rates. The results were then compared to the actual retention rates. The results follow in Table 4-1.

32

Table 4-1: Predicted Retention Rates for FY 94.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
8	88.00%	82.30%	95.88%	94.76%	YES
9	90.14%	85.84%	97.83%	97.71%	YES
10	91.64%	84.97%	98.18%	96.84%	YES
11	93.12%	80.68%	95.79%	95.74%	YES
12	91.87%	81.24%	93.46%	89.67%	YES
13	90.73%	86.38%	95.44%	88.94%	YES
14	96.53%	89.45%	98.57%	94.77%	YES
15	94.93%	90.45%	96.92%	93.63%	YES

Figure 4-3 : Graph of retention rates vs. year of service for 1994

33
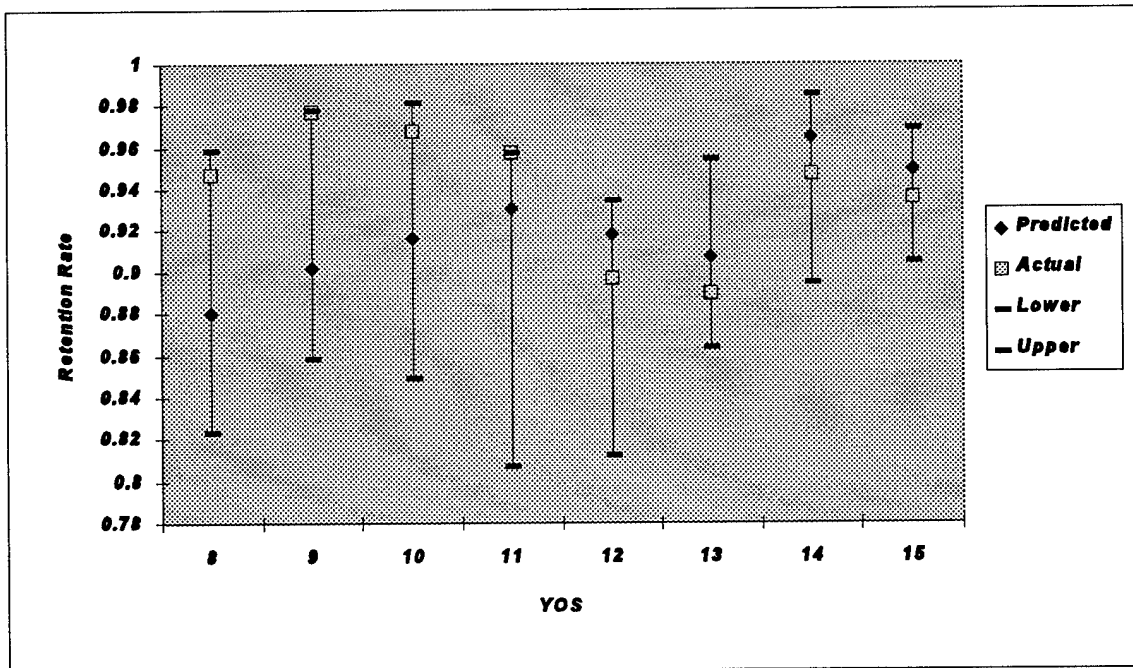
One should notice that all of the actual retention rates fall within the 95% prediction interval.

Model Revision Results

The model was then revised using the FY 94 data along with the original data. Most times when an analyst revises the model, he expects the variables and/or coefficients not to change that drastically. In this effort, the researcher thought that this would not be the case for these models. The reason being that the retention patterns have gone through tremendous change because of the downsizing. The downsizing has forced all of the voluntary exits to leave earlier than seen in previous data. Now at the end of the downsizing process, the retention rates are unnaturally high, because all of those who would historically leave, have already left through exit incentives. This can easily been seen in the 1994 predictions for YOS 8 - 11. One would expect the model revisions to settle down, once all of this downsizing turmoil settles down. This will be addressed once again in the next chapter on conclusions and recommendations.

The ANOVA table for the prediction model selected for Group 1 based on data from 1976 - 1994 is as follows:

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	15	58.42846	3.89523	55.158	0.0001
Error	153	10.80551	0.07062		
C Total	168	69.23397			

R-square   0.8439   Adj R-sq   0.8252

Parameter Estimates

Variable	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
Intercept	2.140962	0.28540935	7.501	0.0001
AADper	1.867857	0.18726005	9.975	0.0001
AADperE	-1.665285	0.38207628	-4.359	0.0001
AADY6B	22.97738	4.50335854	5.102	0.0001
Defpro1Y	60.044899	0.00676611	6.636	0.0001
Y6	-3.065561	0.38531819	-7.956	0.0001
Paydif1Y13E	0.387052	0.08910014	4.344	0.0001
Plt1	-0.000046	0.00001131	-4.060	0.0001
Plt1Y9E	0.000635	0.00014695	4.323	0.0001
Plt1Y10E	0.000473	0.00014273	3.317	0.0011
Urate2	0.044084	0.00901942	4.888	0.0001
Lag2E	0.011541	0.00208967	5.523	0.0001
Plt2Y7	-0.000060	0.00002052	-2.910	0.0042
Plt2Y8	-0.000051	0.00001878	-2.722	0.0073
Irate2Y6B	-0.291213	0.05698227	-5.111	0.0001
Sent3	-0.010416	0.00337926	-3.082	0.0024

Order Variables entered with respective partial R^2 value

Order	Variable	Partial $R^{**}2$	Model $R^{**}2$
1	Plt2Y7	0.1996	0.1996
2	Plt2Y8	0.1558	0.3554
3	AADY6B	0.0666	0.4220
4	Plt1Y9E	0.0569	0.4789
5	Lag2E	0.0554	0.5343
6	Defpro1Y6	0.0444	0.5787
7	Urate2	0.0443	0.6230
8	Irate2Y6B	0.0368	0.6598
9	Y6	0.0367	0.6965
10	Plt1	0.0316	0.7281
11	AADper	0.0303	0.7584
12	Paydif1Y13E	0.0284	0.7868
13	AADperE	0.0207	0.8075
14	Sent3	0.0152	0.8227
15	Plt1Y10E	0.0122	0.8349

35

where Defpro1 :   Defense Production index for the previous year,
       Lag2 :   the lagging economic indicator from two years previous,
       Irate2 :   the federal funds interest rate from two years previous,
       E suffix :   the exit incentive indicator variable,   .
       B suffix :   the pilot bonus indicator variable,
and the rest of the variables are defined above.

It is interesting to note that not only has the number of independent variables
increased by six variables, but the R-squared value has increased to .8439. One would
expect that with adding another year's worth of data, that the R-square value would
remain constant or even decrease. This increase means that the things for Group 1 at
least might be starting to settle down and the percentage of those pilots that have taken
the bonus has increased, thereby making it easier to predict retention rates.

Heteroscedasticity is still not a problem after adding the 1994 data, as seen in
figure 4 - 4 below.

Figure 4-4 : Plot of Studentized Residuals vs. Predicted Value of Logit
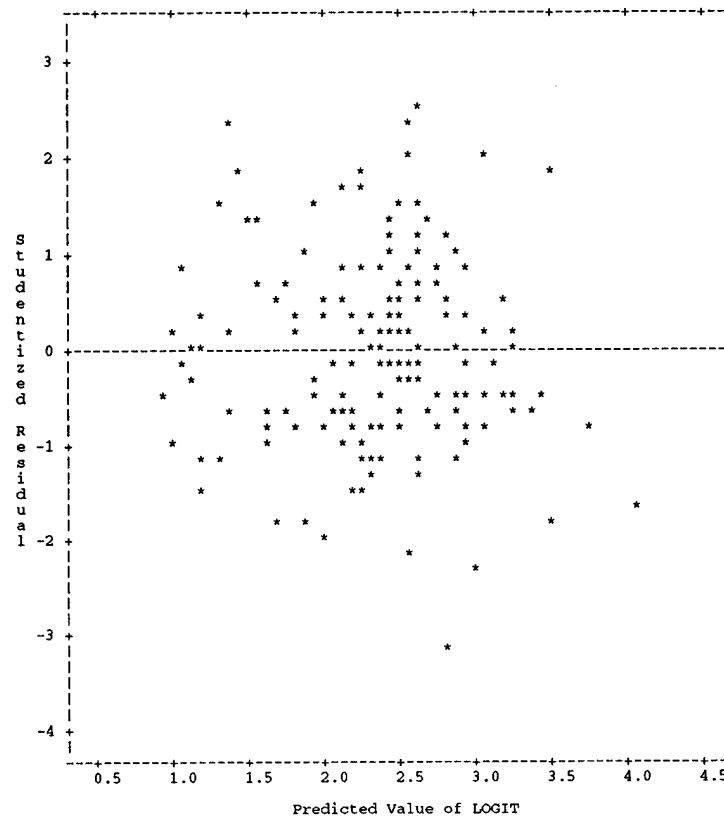
The calculated Durbin-Watson statistic was 2.184. Since this is still very close to 2, the

assumption of independence is still valid. The calculated Wilkes-Shapiro statistic was

0.9853, with a 0.7279 probability of obtaining a value less then this when sampling from

a normal population. This indicates that the residuals are still normal. The normal

probability plot in figure 4 - 5 also indicates this normality.

Figure 4-5 : Normal Probability Plot of the Studentized Residuals
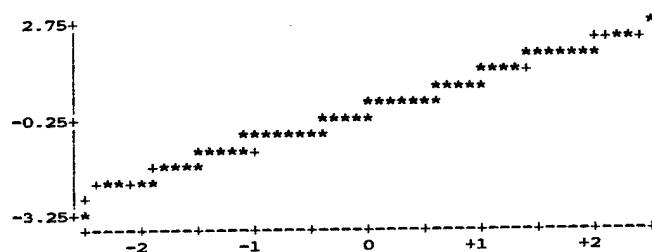
The researcher used the final retention rates for 1995 for a final check on the models and final model revision. Group 1's predicted retention rates for 1995 and 1996 are in tables 4 - 2 and 4 - 3 respectively.

Table 4-2 : Retention rates for Group1 for 1995

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
8	91.51%	85.52%	95.16%	90.16%	YES
9	97.60%	95.11%	98.84%	96.55%	YES
10	96.90%	93.74%	98.50%	98.00%	YES
11	92.69%	87.82%	96.71%	96.41%	YES
12	92.71%	87.85%	95.72%	91.64%	YES
13	95.27%	87.49%	97.42%	89.90%	YES
14	93.11%	88.02%	96.13%	94.33%	YES
15	93.17%	87.97%	96.22%	93.76%	YES

Figure 4-6 : Graph of retention rates vs. year of service for 1995
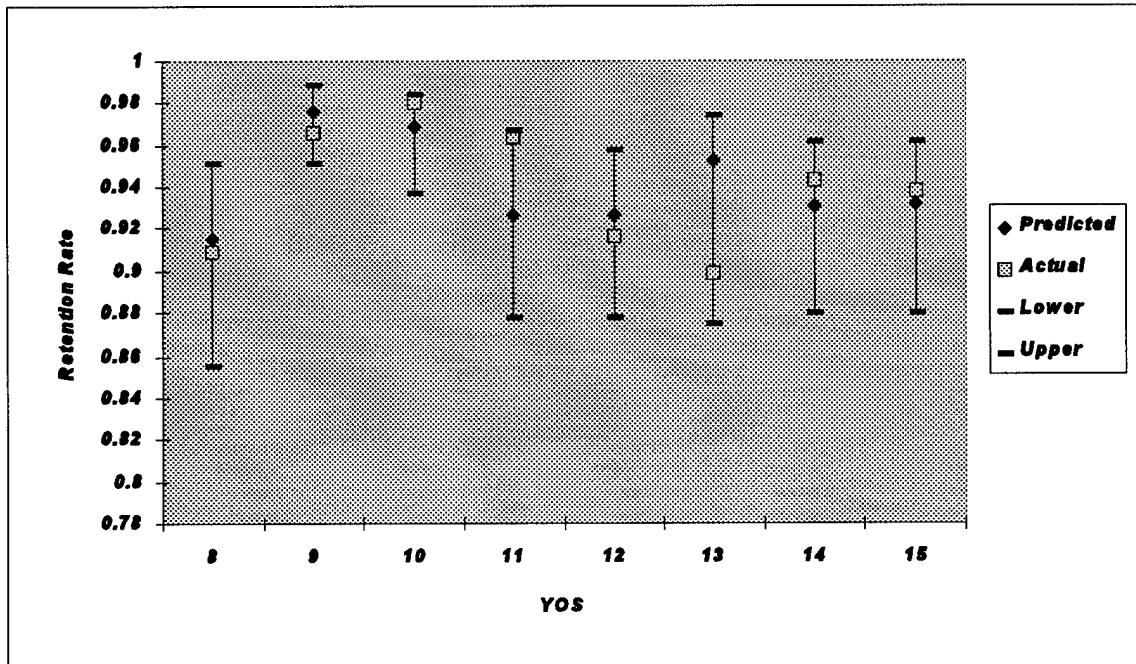
Table 4-3 : Predicted retention rates for Group1 for 1996

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
8	91.17%	84.99%	94.96%
9	98.09%	95.76%	99.15%
10	97.63%	94.19%	98.82%
11	92.49%	87.46%	95.61%
12	92.52%	87.49%	95.63%
13	90.80%	84.73%	94.62%
14	92.77%	87.55%	95.90%
15	92.99%	87.38%	96.22%

V. Conclusions and Recommendations

Introduction

This chapter will summarize the conclusion of the analysis of officer retention rates. The possible applications of these retention models will be discussed. Several recommendations for refining this approach to modeling retention rates will be suggested.

Conclusions

The results that were summarized in the last chapter and in appendix C of this document, truly indicate that the use of the Modified Miller's Method of Variable Selection is a very good analysis tool in the model building process for officer retention rates. For Group 1 (pilots), for instance, the final model explains over 82 % of the total variation of nineteen years worth of retention data. The models for all of the other groups do not explain this much variation. They all do however explain between 45 % to 55 % of the total variation. All of these models, however, are valuable tools for personnel managers, by not only allowing them to accurately predict the retention rates for the next fiscal year, but it also gives them the ability to analyze retention trends by looking at the predictor variables.

The previous chapter has also demonstrated that there is a true statistical relationship (i.e. not random) between the retention rate and certain econometric predictor variables. This means that in most cases, at least half of the reasoning behind

40

an officer's decision to stay or leave is based purely on economics. This is a very good sign to the analyst who is trying to make predictions about retention. It would be nearly impossible to try to predict retention when it is based almost entirely on random reasoning.

Another benefit of these models are there ease of use and maintenance. All of these model are ordinary linear regression models. This technique is widely known and used in the analysis community. All of these models have under 10 predictor variables, which really cuts down on the data requirements for use and refinement.

One of the negative aspects of these models was seen in the model revision in the previous chapter. The model variables and coefficients seem to be changing a lot. This is due to the changing patterns in the retention trends. It would be foolish to believe that one single model will work from now on. Retention patterns are changing now in ways that have never before been seen. Any model built from past data will not be able to pick up these new changes.

Another drawback to these models are that they do not attempt to quantify the impact of the gradual erosion of the retirement benefits and/or healthcare benefits on retention. It seems logical that as the retirement/healthcare benefits are decreased, there would be less and less incentive to stay in the Air Force.

One final drawback to these models is that they show statistical correlation of the predictor variables to the logit transformed retention rates and not directly to the rates themselves. This makes the interpretation of the predictor coefficients a little more difficult. An analyst cannot simply look at the coefficient and say what an increase in

that variable will directly do to retention. However, by using inverse logit transformation, the analyst can indirectly find out what an increase in that variable will do to the retention rate. Therefore, predictor correlation to retention is possible but is not as simple as reading off a coefficient.

Recommendations

Due to the fact that the predictor/coefficients are changing during model revision, the researcher would recommend that these models be revised annually when the next fiscal year data is published. The researcher is optimistic that the changes will cease when the military drawdown and its' effects are over.

One of the main reasons behind the econometric variable changes during model revision is probably due to the fact that most of those econometric variables are highly correlated with each other. One simple remedy to this is by using principle component analysis on the predictor variables. This would not only cut down on the number of possible predictors, but it would also quantify the true underlying essence of those variables. The only drawback to principle components is that interpretation of the predictor variables would become difficult.

If one only wanted to predict officer retention rates, then the Box-Jenkins seasonal auto-regressive integrated moving average (SARIMA) method of model building might be a good avenue for further research. It seems probable that some of the variance in retention rates is based on the retention rates in the past. Therefore, it is suggested to try to use the SARIMA method on the residuals of the method in this effort.

Another suggestion for further research is to investigate the correlation of retirement/healthcare benefits and retention. The major difficulty in this would be how to quantify the value of those benefits now and in the past. Along those same lines, it might be beneficial to investigate the impact of how society views the military. This once again seems very difficult to quantify.

In conclusion, there are many aspects to an officer's decision to stay or leave the Air Force. This research has shown that, in most cases, over half of it is explained by economics. Further research should concentrate on the explanations of some or most of the other half.

Appendix A - Data Description

This appendix will give a short description source of all of the variables in Table 3-1 of this document.

Airline hiring rates: The Future Aviation Professionals of America (FAPA) is the source of this data. This variable is a count of the total number of pilots that are hired in a given year by major, national, and jet airlines. Since most of the pilots leaving the Air Force go and fly for these companies, this variable should be a very good indicator of pilot retention.

Leading Economic Index: The Bureau of Economic Analysis in the U.S. Department of Commerce provides this data in the *Business Conditions Digest* and in the *Survey of Current Business*. This index represents business commitments and expectations regarding labor, product, and financial markets and, thus points to future business actions. The components of this index reflect: the degree of tightness in labor markets due to employer hiring and firing; the buildup of orders, contracts, and inventories that affect future production; materials prices that reflect shortages or gluts of raw materials for which some time will be required to expand or reduce inventories; and financial conditions associated with the availability of funds in credit markets and the optimism and pessimism generated by price movements in the stock market [Frumkin, 1990:164-166].

Coincident Economic Index: The Bureau of Economic Analysis in the U.S. Department of Commerce provides this data in the *Business Conditions Digest* and in the *Survey of Current Business*. This index represents the current level of actual production and sales. The components of this index reflect: employment; real incomes generated from production; output in cyclically sensitive manufacturing and mining industries; and real manufacturing and trades sales depicting the flow of goods from manufacturers to other consuming businesses, as well as to distributors and households [Frumkin, 1990 164-166].

Lagging Economic Index: The Bureau of Economic Analysis in the U.S. Department of Commerce provides this data in the *Business Conditions Digest* and in the *Survey of Current Business*. This index represents whether business costs are rising or falling. The components of this index reflect: the effect of the duration of unemployment on business costs of recruitment and training; the cost of maintaining inventories; labor cost per unit of output; the burden of paying back business and consumer loans; and interest payments as a cost of production [Frumkin, 1990:164-166].

Consumer Sentiment Index: This index is provided by the Survey Research Center of The University of Michigan in the *Surveys of Consumer Attitudes* or in the *Business Conditions Digest*. This index combines three main categories of household attitudes

toward the economy in one figure: (1) expected business conditions in the national economy for one and five years ahead, (2) personal financial well-being compared to one year earlier and expected one year later, and (3) whether the current period is a good or bad time to buy furniture and major household appliances [Frumkin, 1990:53].

Help-wanted Advertising Index: This index is provided by The Conference Board and can be found in the *Business Conditions Digest* and in the *Survey of Current Business*. This index tracks employer's advertisements for job openings in the classified section of newspapers in 51 labor market areas. This index represents job vacancies resulting from turnover in existing positions such as workers changing jobs or retiring and from the creation of new jobs [Frumkin, 1990:123].

Federal Reserve Discount Rate: This variable is provided by the Federal Reserve Board and can be found in many sources. This variable tracks the short-term borrowing by commercial banks from regional Federal Reserve banks to meet seasonal demands for money, to maintain certain reserve levels over a two-week period, to meet huge outflows at the end of a day, or to keep bank reserves from falling close to or below legal minimum requirements [Frumkin, 1990:146].

Employment Cost Index: The Bureau of Labor Statistics in the U.S. Department of Labor provides this index in the *Monthly Labor Review* and *Current Wage Developments*. This index measures the changes in labor costs for money wages and salaries and noncash fringe benefits in nonfarm private industry and state and local governments for workers at all levels of responsibility. Thus, the ECI represents labor costs for the same jobs over time [Frumkin, 1990:84].

Military Pay Increase Rate: This variable can be found in the Armed Services Almanac. This measures the change in salaries for the Armed Services.

Pay Difference: This is simply MPIR minus ECI. Thus, this variables measures the year to year difference in the changes of wages between the military sector and the private sector. The reasoning for using a year to year difference versus a cumulative sum difference is that most officers don't usually track the cumulative difference, but they usually are aware of pay increases on the outside versus their pay increase.

Civilian Unemployment Rate: The Bureau of Labor Statistics in the U.S. Department of Labor provides this index in the *Monthly Labor Review* and in the *Survey of Current Business*. This variable is the percentage that unemployed persons are of the total labor force, and the labor force is defined as the sum of the employed and unemployed. The unemployed is defined as the number of persons without jobs who are available for and actively seeking work. It covers all persons 16 years and older who lost or quit previous jobs as well as school graduates, students, and others with no work experience or who re-enter the workplace [Frumkin, 1990:224].

White Collar Unemployment Rate: The Bureau of Labor Statistics in the U.S. Department of Labor provides this index in the *Employment and Earnings*. This variable is the percentage that unemployed white collar workers are of the total white collar labor force.

Advanced Academic Degree Percentage: This variable was provided by AFMPC Analysis Division. This variable is the percentage that officers with AADs are of the total officers in that particular AFSC and Yeargroup combination.

Pilot Bonus Eligibility: This variable was provided by AFMPC Analysis Division. This variable is a binary variable that indicates whether that particular yeargroup was eligible for the pilot bonus in that particular year.

RIF or VSI/SSB/TERA: These variables were provided by AFMPC Analysis Division. These are two binary variables to indicate whether a RIF occurred in that year, or whether there were exit bonuses that year. The reasoning behind one variable for all three exit bonus programs is that all of these programs happened together in the same years. Therefore, if three variables were used, they would be linear combinations (perfect copies) of each other.

Appendix B - Models for AFSC's

This appendix will outline the ANOVA tables for each of the separate AFSC's built with data from FY 76 - 93.

AFSC 11 (pilots)

R-square = 0.73185334

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	17	99.39547967	5.84679292	33.71	0.0001
Error	210	36.41790512	0.17341860		
Total	227	135.81338479			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.24025586	0.29520378	9.98727670	57.59	0.0001
AADPER	-1.23325841	0.15805361	10.55832471	60.88	0.0001
AADY6	-9.34259040	1.14774100	11.49059436	66.26	0.0001
AADY9	-1.20295391	0.44477684	1.26855502	7.31	0.0074
Sent1	0.01098767	0.00326854	1.95973762	11.30	0.0009
Irate1	0.04976766	0.00933028	4.93402140	28.45	0.0001
Pilot1Y5	0.00008577	0.00002668	1.79225498	10.33	0.0015
Pilot1Y6	0.00012990	0.00003263	2.74861873	15.85	0.0001
Pilot1Y10	-0.00015960	0.00002291	8.41326984	48.51	0.0001
Pilot1Y11	-0.00014251	0.00002287	6.73415493	38.83	0.0001
Pilot1Y12	-0.00009657	0.00002300	3.05740489	17.63	0.0001
Lag2Y8	-0.00761785	0.00169794	3.49073132	20.13	0.0001
Paydif2Y6	-0.10135839	0.03986923	1.12082872	6.46	0.0117
Pilot2Y8	-0.00017739	0.00003539	4.35763137	25.13	0.0001
Pilot2Y9	-0.00014111	0.00003384	3.01540218	17.39	0.0001
Sent3Y7	-0.02115624	0.00139084	40.12558782	231.38	0.0001
Paydif3Y5	-0.18112528	0.03726163	4.09760540	23.63	0.0001
Paydif3Y6	-0.08756400	0.04237507	0.74050147	4.27	0.0400

AFSC 12 (Navigators)

R-square = 0.44751653

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	8	13.61889327	1.70236166	19.14	0.0001
Error	189	16.81326346	0.08895907		
Total	197	30.43215673			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	1.70710327	0.19687960	6.68818609	75.18	0.0001
Coin1	0.00621858	0.00209580	0.78319989	8.80	0.0034
Irate1Y6	-0.05422169	0.00818528	3.90363076	43.88	0.0001
Sent1Y13E	-0.00949467	0.00267612	1.11979927	12.59	0.0005
Irate1Y8E	0.78523670	0.22729539	1.06172030	11.93	0.0007
Paydif1Y7E	0.37070129	0.09512791	1.35089774	15.19	0.0001
Paydif1Y8E	1.56340854	0.33089062	1.98594066	22.32	0.0001
Lag2Y10E	0.01500585	0.00309357	2.09310606	23.53	0.0001
Sent2Y5E	0.01415038	0.00389398	1.17473478	13.21	0.0004

AFSC 13 (Space, Missile, ATC)

R-square = 0.45599831

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	72.27738377	12.04623063	30.87	0.0001
Error	221	86.22623671	0.39016397		
Total	227	158.50362047			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.85637956	0.05541842	1036.50322795	2656.58	0.0001
PAYDIF2	-0.06407140	0.01662602	5.79427607	14.85	0.0002
Irate1Y14	0.10722728	0.01701000	15.50418135	39.74	0.0001
Paydif1E	0.44760718	0.06018948	21.57744651	55.30	0.0001
Urate3Y9	0.08525351	0.02150948	6.12931554	15.71	0.0001
Irate3Y15	0.12522602	0.01613705	23.49568065	60.22	0.0001
Paydif3Y7E	-1.30429530	0.38533479	4.47015505	11.46	0.0008

AFSC 14 (Intelligence)

R-square = 0.33855816

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	39.19045432	7.83809086	22.73	0.0001
Error	222	76.56647738	0.34489404		
Total	227	115.75693169			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.67317378	0.04964957	999.79066970	2898.83	0.0001
Help1Y11	-0.00319494	0.00114604	2.68046643	7.77	0.0058
Paydif1R	0.41200345	0.05313734	20.73422519	60.12	0.0001
Urate2Y9	0.10022017	0.02002076	8.64241435	25.06	0.0001
Help2Y14	0.00317177	0.00114228	2.65917389	7.71	0.0060
Defpro2Y15	0.00719352	0.00186969	5.10539441	14.80	0.0002

AFSC 15 (Weather)

R-square = 0.39932550

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	7	60.78622441	8.68374634	20.89	0.0001
Error	220	91.43602324	0.41561829		
Total	227	152.22224765			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.12083852	0.12223570	125.11625790	301.04	0.0001
AADPER	0.70381574	0.19202730	5.58324441	13.43	0.0003
Irate1Y14	0.06990743	0.01827409	6.08232603	14.63	0.0002
Irate2E	-0.10605090	0.01929452	12.55612028	30.21	0.0001
Urate3Y8	0.12402294	0.02232685	12.82459600	30.86	0.0001
White3Y9	0.33611660	0.05965373	13.19469652	31.75	0.0001
Irate3Y6	0.07995810	0.01740648	8.76995907	21.10	0.0001
Paydif3Y6E	-1.65361178	0.39690489	7.21421135	17.36	0.0001

AFSC 21 (Maintenance)

R-square = 0.53761404

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	9	74.87975275	8.31997253	28.16	0.0001
Error	218	64.40186301	0.29542139		
Total	227	139.28161576			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.42671380	0.05209111	641.13852515	2170.25	0.0001
AADY15	1.13845113	0.17996633	11.82192517	40.02	0.0001
Help1Y7	0.00640893	0.00107620	10.47674355	35.46	0.0001
Help1Y8	0.01030433	0.00112330	24.85933165	84.15	0.0001
Irate1Y9	0.09138312	0.01507648	10.85361064	36.74	0.0001
Irate1Y14	0.05314765	0.01628293	3.14735053	10.65	0.0013
Paydif1R	0.35806188	0.05119901	14.44890588	48.91	0.0001
Help2Y6	0.00499068	0.00107257	6.39598218	21.65	0.0001
Paydif3Y14	0.15236783	0.04953605	2.79502538	9.46	0.0024
Irate3Y8E	-0.15886870	0.04363963	3.91522349	13.25	0.0003

AFSC 23 (Supply)

R-square = 0.35044134

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	8	71.72192272	8.96524034	14.77	0.0001
Error	219	132.93978369	0.60703098		
Total	227	204.66170641			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.58602871	0.07396693	741.99736711	1222.34	0.0001
AADperE	-0.68927136	0.41953936	1.63849794	2.70	0.1018
Defpro1Y12	-0.00436108	0.00247674	1.88208132	3.10	0.0797
Paydif1E	0.24553666	0.13842694	1.90986346	3.15	0.0775
Help2Y8	0.00568026	0.00153721	8.28859106	13.65	0.0003
Help2Y14	0.00567700	0.00154375	8.20911591	13.52	0.0003
Irate2Y7	0.09127196	0.02090930	11.56662574	19.05	0.0001
Irate3Y9	0.08860988	0.02052079	11.31845443	18.65	0.0001
Irate3Y15	0.09734029	0.02058163	13.57802554	22.37	0.0001

AFSC 24 (Transportation)

R-square = 0.28276558

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	51.85543339	8.64257223	14.52	0.0001
Error	221	131.53122076	0.59516389		
Total	227	183.38665415			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.64048823	0.06534708	971.74572578	1632.74	0.0001
Sent1Y12	-0.00733418	0.00226001	6.26786261	10.53	0.0014
Irate1Y7	0.08470252	0.02109942	9.59153692	16.12	0.0001
Paydif1E	0.40759767	0.07295056	18.57987394	31.22	0.0001
Paydif1Y15E	-0.63979676	0.26184789	3.55322330	5.97	0.0153
Defpro2Y15	0.00538648	0.00273492	2.30864031	3.88	0.0501
Irate3Y9	0.07434669	0.02002914	8.20040650	13.78	0.0003

AFSC 25 (Logistics)

R-square = 0.27475613

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	53.04914499	17.68304833	28.29	0.0001
Error	224	140.02806095	0.62512527		
Total	227	193.07720594			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.73858353	0.06098144	1260.73460898	2016.77	0.0001
Help1Y9	0.00600017	0.00152720	9.64939726	15.44	0.0001
Irate1Y15	0.08788740	0.02143523	10.50906145	16.81	0.0001
Paydif1E	0.50635421	0.07162359	31.24379059	49.98	0.0001

AFSC 31 (Security Police)

R-square = 0.22751034

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	42.95850931	14.31950310	21.99	0.0001
Error	224	145.86151794	0.65116749		
Total	227	188.82002725			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.99726433	0.06210387	1516.72011813	2329.23	0.0001
Irate1Y10	-0.08263494	0.02187208	9.29479802	14.27	0.0002
Paydif1E	0.46937174	0.07312875	26.82564929	41.20	0.0001
Irate2Y11	-0.08796367	0.02116405	11.24869161	17.27	0.0001

AFSC 32 (Civil Engineers)

R-square = 0.29712264

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	3	47.41992210	15.80664070	31.56	0.0001
Error	224	112.17721405	0.50079113		
Total	227	159.59713614			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	1.74694342	0.14252930	75.23258705	150.23	0.0001
AADper	1.59649679	0.22333017	25.59162898	51.10	0.0001
Defpro1Y12	-0.00765419	0.00224519	5.82035695	11.62	0.0008
Paydif1E	0.40168504	0.06398063	19.73926042	39.42	0.0001

AFSC 33 (Communications-Computer Systems)

R-square = 0.46157491

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	8	47.50270308	5.93783788	23.47	0.0001
Error	219	55.41169242	0.25302143		
Total	227	102.91439550			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.06239053	0.10368777	100.10229164	395.63	0.0001
AADper	1.30165292	0.17530997	13.94872275	55.13	0.0001
Paydif1Y13	-0.15104590	0.04484952	2.86985237	11.34	0.0009
Paydif1E	0.35073409	0.04687906	14.16300908	55.98	0.0001
Help2Y10	-0.00510153	0.00099976	6.58825475	26.04	0.0001
Help3Y11	-0.00585021	0.00100205	8.62428734	34.09	0.0001
Sent3Y12	-0.00913353	0.00156878	8.57645309	33.90	0.0001
Irate3Y9	0.03888797	0.01317248	2.20522314	8.72	0.0035
White3Y13E	-0.58488157	0.13635116	4.65559799	18.40	0.0001

AFSC 37 (Information Management)

R-square = 0.37003183

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	52.71566200	8.78594367	21.64	0.0001
Error	221	89.74684426	0.40609432		
Total	227	142.46250627			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.19495097	0.13701257	104.22111739	256.64	0.0001
AADper	1.02356561	0.26740966	5.94982365	14.65	0.0002
Sent1X12	-0.01089945	0.00195255	12.65411212	31.16	0.0001
Irate1X9	0.06595959	0.01700700	5.76447094	14.19	0.0002
Defpro1X13	-0.01009060	0.00211794	9.21797272	22.70	0.0001
Paydif1R	0.42696818	0.05798315	22.01984922	54.22	0.0001
Help2X11	-0.00445341	0.00126970	4.99584719	12.30	0.0005

AFSC 61 (Scientists)

R-square = 0.38048463

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	58.51622581	11.70324516	27.27	0.0001
Error	222	95.27770148	0.42917884		
Total	227	153.79392729			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	-4.91931367	1.30657723	6.08383117	14.18	0.0002
AADper	2.90631790	0.33447825	32.40326352	75.50	0.0001
AADperE	-3.48085007	0.54563200	17.46663558	40.70	0.0001
Sent1Y14	0.00761019	0.00194670	6.55889373	15.28	0.0001
Lag3	0.05497033	0.01315204	7.49736474	17.47	0.0001
White3E	0.87788880	0.18832343	9.32628304	21.73	0.0001

AFSC 62 (Engineers)

R-square = 0.41874678

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	7	49.40959167	7.05851310	22.64	0.0001
Error	220	68.58436989	0.31174714		
Total	227	117.99396156			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	0.82450937	0.20397312	5.09386894	16.34	0.0001
AADper	1.87591104	0.22683482	21.32099688	68.39	0.0001
AADperE	-0.61549493	0.14078321	5.95867500	19.11	0.0001
AADperY7	-2.94489696	1.11685759	2.16743999	6.95	0.0090
Irate1Y8	0.05514635	0.01522430	4.09036066	13.12	0.0004
Urate2Y7	0.34603229	0.09684863	3.97968543	12.77	0.0004
Irate2Y9	0.07025944	0.01475073	7.07268809	22.69	0.0001
Irate3	0.04896899	0.01337337	4.17987411	13.41	0.0003

AFSC 63 (Acquisition)

R-square = 0.38970305

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	7	56.06715107	8.00959301	20.07	0.0001
Error	220	87.80431998	0.39911055		
Total	227	143.87147104			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	-0.49442295	0.45003867	0.48171568	1.21	0.2731
AADper	1.74404812	0.23302265	22.35701651	56.02	0.0001
Urate1	0.12412371	0.05308292	2.18219238	5.47	0.0203
Defpro1	0.01121086	0.00255875	7.66154449	19.20	0.0001
Paydif1E	0.24137475	0.06925652	4.84792079	12.15	0.0006
Irate2	0.03332634	0.02024936	1.08105051	2.71	0.1012
Irate2Y10	-0.05374015	0.01679571	4.08595976	10.24	0.0016
Urate3Y15	0.07867234	0.02263235	4.82256271	12.08	0.0006

AFSC 64 (Contracting)

R-square = 0.28019654

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	6	41.57398547	6.92899758	14.34	0.0001
Error	221	106.80038850	0.48325968		
Total	227	148.37437397			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	2.03307881	0.19064680	54.95787433	113.72	0.0001
AADper	1.17635773	0.26444535	9.56285341	19.79	0.0001
Coin1Y12	-0.00646698	0.00184560	5.93345591	12.28	0.0006
White1Y11	-0.23927048	0.06291787	6.98893203	14.46	0.0002
Irate1Y7	0.05939557	0.01896771	4.73869611	9.81	0.0020
Paydif1E	0.29718518	0.06330886	10.64892651	22.04	0.0001
Paydif2Y13	0.19402628	0.05712411	5.57522880	11.54	0.0008

AFSC 65 (Financial Management)

R-square = 0.30991171

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	45.49116214	9.09823243	19.94	0.0001
Error	222	101.29632720	0.45628976		
Total	227	146.78748935			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEPT	1.96030871	0.16235202	66.52337367	145.79	0.0001
AADper	0.83329208	0.23521053	5.72692774	12.55	0.0005
Help1Y9	0.00573527	0.00130761	8.77792006	19.24	0.0001
Paydif1E	0.34496022	0.06142669	14.39012357	31.54	0.0001
Help2Y8	0.00662149	0.00130246	11.79299396	25.85	0.0001
Paydif2Y14	0.17617253	0.05540545	4.61330555	10.11	0.0017

Appendix C - Models for Groups

This appendix will outline the ANOVA tables for each of the groups, except group1, a list of the order that the variables eneterd, and then give the predictions for FY 95 & 96.

Group 2
The ANOVA table for Group2, based on FY 76 - 94 is as follows:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	6	50.40269	8.40045	29.887	0.0001
Error	209	58.74387	0.28107		
C Total	215	109.14656			

Root MSE	0.53016	R-square	0.4618
Dep Mean	2.82976	Adj R-sq	0.4463
C.V.	18.73521		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEPT	1	2.668013	0.04522653	58.992	0.0001
Irate1Y8	1	0.053190	0.01466628	3.627	0.0004
Irate1Y14	1	0.098450	0.01466628	6.713	0.0001
Defpro1Y15	1	0.011462	0.00167705	6.835	0.0001
Paydif1E	1	0.392191	0.05143780	7.625	0.0001
Plt1Y7E	1	0.000920	0.00026662	3.451	0.0007
Urate2Y9	1	0.110311	0.01836442	6.007	0.0001

Step	Variable Entered Removed	Number In	Partial R**2	Model R**2	C(p)	F	Prob>F
1	Paydif1E	1	0.1533	0.1533	116.7855	38.7522	0.0001
2	Defpro1Y15	2	0.0769	0.2302	88.9150	21.2850	0.0001
3	Irate1Y14	3	0.0881	0.3183	56.7096	27.3944	0.0001
4	Urate2Y9	4	0.0802	0.3986	27.5507	28.1502	0.0001
5	Irate1Y8	5	0.0326	0.4311	16.9099	12.0166	0.0006
6	Plt1Y7E	6	0.0307	0.4618	7.0000	11.9099	0.0007

Predicted Retention Rates for FY 95.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
4	95.69%	88.51%	98.46%	95.75%	YES
5	95.69%	88.51%	98.46%	92.46%	YES
6	95.69%	88.51%	98.46%	96.42%	YES
7	94.26%	90.82%	99.83%	91.39%	YES
8	91.42%	87.33%	98.73%	87.53%	YES

9	93.92%	88.10%	99.28%	89.43%	YES
10	95.69%	84.51%	98.46%	87.00%	YES
11	95.69%	84.51%	98.46%	85.46%	YES
12	95.69%	84.51%	98.46%	85.12%	YES
13	95.69%	86.51%	98.46%	87.88%	YES
14	96.95%	91.68%	98.92%	92.34%	YES
15	92.08%	88.57%	99.34%	89.97%	YES

Predicted Retention Rates for FY 96.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
4	91.93%	79.95%	97.02%
5	91.93%	79.95%	97.02%
6	91.93%	79.95%	97.02%
7	99.03%	95.27%	99.81%
8	93.91%	84.26%	97.80%
9	95.81%	88.72%	98.52%
10	91.93%	79.95%	97.02%
11	91.93%	79.95%	97.02%
12	91.93%	79.95%	97.02%
13	91.93%	79.95%	97.02%
14	95.23%	87.39%	98.29%
15	96.10%	89.46%	98.62%

Group 3

The ANOVA table for Group3, based on FY 76 - 94 is as follows:

```
                        Analysis of Variance

                            Sum of        Mean
        Source      DF      Squares      Square     F Value      Prob>F

        Model        7     68.86830     9.83833      33.261      0.0001
        Error      208     61.52391     0.29579
        C Total    215    130.39221

            Root MSE        0.54386    R-square      0.5282
            Dep Mean        2.82078    Adj R-sq      0.5123
            C.V.           19.28063

                        Parameter Estimates

                      Parameter      Standard     T for H0:
        Variable  DF   Estimate         Error   Parameter=0    Prob > |T|

        INTERCEPT  1   2.546743      0.05161385      49.342      0.0001
        Help1Y6    1   0.004205      0.00108500       3.876      0.0001
        Irate1Y14  1   0.077476      0.01521282       5.093      0.0001
        Irate1Y7   1   0.080293      0.01521282       5.278      0.0001
        Irate1Y9   1   0.096715      0.01521282       6.357      0.0001
        Paydif1E   1   0.396239      0.05257897       7.536      0.0001
        Plt1Y8     1   0.000189      0.00002912       6.479      0.0001
        Irate2Y15  1   0.123768      0.01508446       8.205      0.0001

            Variable         Number   Partial   Model
     Step   Entered Removed    In      R**2      R**2      C(p)        F      Prob>F

       1    Paydif1E           1      0.1986    0.1986   141.2711   53.0402   0.0001
       2    Irate2Y15          2      0.0879    0.2865   104.5302   26.2354   0.0001
       3    Plt1Y8             3      0.0558    0.3423    81.9167   17.9984   0.0001
       4    Irate1Y9           4      0.0588    0.4011    57.9954   20.7178   0.0001
       5    Irate1Y7           5      0.0437    0.4448    40.7437   16.5188   0.0001
       6    Irate1Y14          6      0.0493    0.4941    21.0215   20.3565   0.0001
       7    Help1Y6            7      0.0341    0.5282     8.0000   15.0215   0.0001
```

Predicted Retention Rates for FY 95.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
4	95.18%	86.94%	98.32%	98.26%	YES
5	95.18%	86.94%	98.32%	92.27%	YES
6	96.97%	91.39%	98.98%	99.23%	YES
7	96.36%	89.92%	98.74%	93.46%	YES
8	96.57%	90.58%	98.83%	92.86%	YES
9	95.18%	86.94%	98.32%	87.58%	YES
10	95.18%	86.94%	98.32%	90.84%	YES
11	95.18%	86.94%	98.32%	87.14%	YES

12	95.18%	86.94%	98.32%	87.02%	YES
13	95.18%	86.94%	98.32%	87.71%	YES
14	96.32%	89.83%	98.73%	90.91%	YES
15	96.63%	90.63%	98.84%	90.43%	YES

Predicted Retention Rates for FY 96.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
4	90.96%	77.42%	96.73%
5	90.96%	77.42%	96.73%
6	94.59%	85.28%	98.14%
7	94.08%	84.30%	97.92%
8	94.04%	84.25%	97.90%
9	94.58%	85.50%	98.10%
10	90.96%	77.42%	96.73%
11	90.96%	77.42%	96.73%
12	90.96%	77.42%	96.73%
13	90.96%	77.42%	96.73%
14	93.99%	84.09%	97.89%
15	94.06%	84.32%	97.80%

Group 4

The ANOVA table for Group4, based on FY 76 - 94 is as follows:

```
                          Analysis of Variance

                            Sum of        Mean
        Source      DF      Squares      Square     F Value     Prob>F

        Model        8     44.47684     5.55961     30.846      0.0001
        Error      207     37.30868     0.18024
        C Total    215     81.78552

            Root MSE        0.42454    R-square      0.5438
            Dep Mean        2.60128    Adj R-sq      0.5262
            C.V.           16.32050

                          Parameter Estimates

                     Parameter      Standard     T for H0:
        Variable  DF   Estimate        Error     Parameter=0    Prob > |T|

        INTERCEPT  1   2.317883     0.04408159      52.582        0.0001
        Help1Y7    1   0.004762     0.00085782       5.552        0.0001
        Irate1Y8   1   0.078001     0.01201545       6.492        0.0001
        Irate1Y9   1   0.098495     0.01201545       8.197        0.0001
        Irate1Y13  1   0.040091     0.01201545       3.337        0.0010
        Paydif1E   1   0.311470     0.04106165       7.585        0.0001
        Help2Y14   1   0.005384     0.00086322       6.237        0.0001
        Irate2Y6   1   0.051651     0.01191543       4.335        0.0001
        Irate2Y15  1   0.092339     0.01191543       7.750        0.0001

              Variable       Number   Partial   Model
      Step   Entered Removed    In     R**2      R**2      C(p)         F      Prob>F

        1    Paydif1E          1      0.1964    0.1964   152.6400   52.3093    0.0001
        2    Irate1Y9          2      0.0702    0.2667   122.7699   20.3995    0.0001
        3    Irate2Y15         3      0.0694    0.3361    93.2756   22.1617    0.0001
        4    Irate1Y8          4      0.0489    0.3850    73.0650   16.7934    0.0001
        5    Help2Y14          5      0.0523    0.4373    51.3439   19.5088    0.0001
        6    Help1Y7           6      0.0482    0.4855    31.4538   19.5971    0.0001
        7    Irate2Y6          7      0.0338    0.5193    18.1328   14.6093    0.0002
        8    Irate1Y13         8      0.0245    0.5438     9.0000   11.1328    0.0010
```

Predicted Retention Rates for FY 95.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
4	93.47%	85.95%	97.10%	93.69%	YES
5	93.47%	85.95%	97.10%	91.14%	YES
6	94.36%	87.74%	97.50%	93.80%	YES
7	93.12%	89.26%	98.32%	89.40%	YES
8	92.01%	85.05%	97.80%	86.07%	YES
9	91.35%	84.76%	97.96%	85.63%	YES
10	91.47%	84.95%	97.10%	85.54%	YES

11	93.47%	85.95%	97.10%	87.63%	YES
12	93.47%	80.95%	97.10%	82.02%	YES
13	94.31%	81.63%	97.48%	82.15%	YES
14	96.10%	90.26%	98.31%	90.93%	YES
15	94.98%	89.00%	97.79%	89.77%	YES

Predicted Retention Rates for FY 96.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
4	89.39%	78.41%	95.13%
5	89.39%	78.41%	95.13%
6	91.05%	81.41%	95.94%
7	94.03%	86.92%	97.39%
8	92.93%	84.92%	96.84%
9	93.66%	86.35%	97.18%
10	89.39%	78.41%	95.13%
11	89.39%	78.41%	95.13%
12	89.39%	78.41%	95.13%
13	91.37%	81.93%	96.11%
14	94.00%	86.92%	97.36%
15	92.19%	83.56%	96.48%

Group 5

The ANOVA table for Group5, based on FY 76 - 94 is as follows:

```
                           Analysis of Variance

                                 Sum of          Mean
          Source        DF      Squares        Square     F Value     Prob>F

          Model         10     54.58955       5.45895      23.808     0.0001
          Error        193     44.25297       0.22929
          C Total      203     98.84251

              Root MSE          0.47884      R-square       0.5523
              Dep Mean          2.68860      Adj R-sq       0.5291
              C.V.             17.81011

                           Parameter Estimates

                       Parameter        Standard      T for H0:
          Variable  DF   Estimate          Error    Parameter=0     Prob > |T|

          INTERCEPT  1   0.547411     0.25217459          2.171        0.0312
          AADper     1   1.532053     0.21281811          7.199        0.0001
          Urate1     1  -0.212399     0.05486629         -3.871        0.0001
          Sent1Y15   1   0.008439     0.00158035          5.340        0.0001
          Irate1Y9   1   0.058174     0.01363274          4.267        0.0001
          Urate2     1   0.210233     0.04213029          4.990        0.0001
          Urate2Y7   1   0.079772     0.01764469          4.521        0.0001
          Irate2     1   0.095862     0.01563310          6.132        0.0001
          Irate2Y6   1   0.035741     0.01394551          2.563        0.0111
          Plt2Y14    1   0.000095     0.00002705          3.510        0.0006
          Irate3Y8   1   0.068973     0.01342337          5.138        0.0001

                 Variable         Number   Partial    Model
          Step   Entered Removed    In      R**2      R**2       C(p)           F     Prob>F

            1    AADper             1      0.2078    0.2078    141.4980     52.9893    0.0001
            2    Irate2             2      0.1007    0.3085    100.1083     29.2555    0.0001
            3    Sent1Y15           3      0.0366    0.3450     86.3503     11.1620    0.0010
            4    Irate3Y8           4      0.0371    0.3821     72.3725     11.9366    0.0007
            5    Urate2Y7           5      0.0339    0.4160     59.7490     11.5013    0.0008
            6    Irate1Y9           6      0.0338    0.4498     47.1975     12.0855    0.0006
            7    Plt2Y14            7      0.0291    0.4789     36.6437     10.9531    0.0011
            8    Urate2             8      0.0236    0.5024     28.4888      9.2321    0.0027
            9    Urate1             9      0.0346    0.5371     15.5684     14.5042    0.0002
           10    Irate2Y6          10      0.0152    0.5523     11.0000      6.5684    0.0111
```

Predicted Retention Rates for FY 95.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
4	80.50%	61.02%	96.59%	95.16%	YES
5	81.61%	62.80%	97.11%	94.68%	YES
6	86.98%	71.89%	94.58%	92.95%	YES
7	91.86%	80.91%	96.78%	91.40%	YES
8	90.62%	78.71%	96.19%	84.26%	YES
9	90.74%	78.95%	96.25%	88.68%	YES

10	89.76%	77.07%	95.81%	86.83%	YES
11	90.40%	78.28%	96.09%	86.84%	YES
12	90.59%	78.66%	96.18%	83.95%	YES
13	91.09%	79.63%	96.40%	84.95%	YES
14	92.77%	83.04%	97.11%	92.46%	YES
15	95.69%	89.24%	98.35%	93.46%	YES

Predicted Retention Rates for FY 96.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
4	80.91%	61.55%	91.79%
5	85.05%	68.58%	93.68%
6	87.30%	72.49%	94.71%
7	92.35%	82.04%	96.96%
8	90.99%	79.53%	96.33%
9	92.73%	82.96%	97.09%
10	90.88%	79.31%	96.28%
11	91.60%	80.71%	96.60%
12	91.73%	80.97%	96.66%
13	91.72%	80.96%	96.65%
14	93.41%	84.45%	97.37%
15	96.47%	91.09%	98.65%

AFSC 12 (Navigators):

The ANOVA table for AFSC 12, based on FY 76 - 94 is as follows:

```
                    Analysis of Variance

                          Sum of       Mean
         Source      DF   Squares      Square      F Value     Prob>F

         Model        8   13.61889     1.70236     19.136      0.0001
         Error      189   16.81326     0.08896
         C Total    197   30.43216

              Root MSE          0.29826     R-square    0.4475
              Dep Mean          2.23964     Adj R-sq    0.4241
              C.V.             13.31732

                      Parameter Estimates

                          Parameter    Standard     T for H0:
         Variable    DF    Estimate      Error     Parameter=0   Prob > |T|

         INTERCEPT    1    1.707103    0.19687960    8.671       0.0001
         Coin1        1    0.006219    0.00209580    2.967       0.0034
         Irate1Y6     1   -0.054222    0.00818528   -6.624       0.0001
         Sent1Y13E    1   -0.009495    0.00267612   -3.548       0.0005
         Irate1Y8E    1    0.785237    0.22729539    3.455       0.0007
         Paydif1Y7E   1    0.370701    0.09512791    3.897       0.0001
         Paydif1Y8E   1    1.563409    0.33089062    4.725       0.0001
         Lag2Y10E     1    0.015006    0.00309357    4.851       0.0001
         Sent2Y5E     1    0.014150    0.00389398    3.634       0.0004

              Variable        Number   Partial    Model
        Step  Entered Removed    In      R**2      R**2     C(p)         F      Prob>F

         1    Irate1Y6          1      0.1290    0.1290  103.9479    29.0392    0.0001
         2    Lag2Y10E          2      0.0811    0.2102   78.2001    20.0253    0.0001
         3    Paydif1Y8E        3      0.0595    0.2697   59.8441    15.8061    0.0001
         4    Sent2Y5E          4      0.0475    0.3172   45.5838    13.4351    0.0003
         5    Irate1Y8E         5      0.0396    0.3568   34.0451    11.8132    0.0007
         6    Paydif1Y7E        6      0.0360    0.3927   23.7427    11.3109    0.0009
         7    Sent1Y13E         7      0.0291    0.4218   15.8040     9.5466    0.0023
         8    Coin1             8      0.0257    0.4475    9.0000     8.8040    0.0034
```

Predicted Retention Rates for FY 95.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound	Actual Retention Rate	In Bounds
5	97.29%	93.80%	99.84%	99.17%	YES
6	90.11%	83.57%	94.30%	93.98%	YES
7	94.35%	89.85%	96.92%	95.24%	YES
8	96.91%	92.00%	99.99%	95.38%	YES
9	91.74%	85.96%	96.27%	94.14%	YES
10	97.92%	95.37%	99.08%	94.39%	YES

11	91.74%	85.96%	96.27%	96.18%	YES
12	91.74%	85.96%	96.27%	86.79%	YES
13	82.53%	69.03%	90.92%	84.38%	YES
14	91.74%	85.96%	96.27%	94.80%	YES
15	91.74%	85.96%	96.27%	91.83%	YES

Predicted Retention Rates for FY 96.

YOS	Predicted Rate	95% Lower Bound	95% Upper Bound
5	97.61%	94.30%	99.02%
6	89.35%	82.10%	93.88%
7	90.15%	83.31%	94.37%
8	99.75%	97.63%	99.97%
9	91.95%	86.27%	95.41%
10	97.99%	95.51%	99.12%
11	91.95%	86.27%	95.41%
12	91.95%	86.27%	95.41%
13	82.52%	68.78%	91.00%
14	91.95%	86.27%	95.41%
15	91.95%	86.27%	95.41%

Bibliography

Cox, D. R. The Analysis of Binary Data. London: Methuen and CO LTD, 1970.

Cromer, DeJuan and Mark R. Julicher. An Examination of the Effects of Economic Conditions on Pilot Retention. MS thesis, AFIT/LSSR 38-82, School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1982. (AD-A122980)

DeVany, Arthur S. and others. Supply Rate and Equilibrium Inventory of Air Force Enlisted Personnel: A Simultaneous Model of the Accession and Retention Markets Incorporating Force Level Constraints. Air Force Human Resources Laboratory (AFSC), Brooks AFB TX, May 1978. (AD-A058097)

Frumkin, Norman. Guide to Economic Indicators. New York: M.E. Sharpe, Inc, 1990.

Gotz, Glenn and John J. McCall. The Dynamic Retention Model for Air Force Officers, Theory and Estimates. R-3028-AF. The RAND Corp, Santa Monica CA, December 1984.

Guzowski, Bruce A. A Methodology for Long-Term Forecasts of Air Force Pilot Retention Rates: A Management Perspective. MS thesis, AFIT/GSM/LSR/90S-11, School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1990. (AD-A229541)

Hansen, Ross J. A Comparison of Variable Selection Criteria for Multiple Regression: A Simulation Study. MS thesis, AFIT/GOR-88D-3, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1988.

Hosmer, David W. and Stanley Lemeshow. Applied Logistic Regression. New York: John Wiley and Sons, 1989.

Makridakis, Spyros and others. Forecasting: Methods and Applications. New York: John Wiley and Sons, 1983.

Miller, Alan J. "Selection of Subsets of Regression Variables", Journal of Royal Statistical Society, Series A, 147, part 3: 389-425 (1984).

Miller, Alan J. Subset Selection in Regression. Melbourne: Chapman and Hall, 1990.

Mutlu, Ertem. A Comparison of Variable Selection Criteria for Multiple Regression: A Third Simulation Study. MS thesis, AFIT/GOR-94M, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 1994. (AD-A278678)

Neter, John and others. Applied Linear Statistical Models. Boston: Irwin, 1990.

Roth, Russell Theodore. The Determinants of Career Decisions of Air Force Pilots. PhD dissertation. Massachusetts Institute of Technology, May 1981. (AD-A107265)

Saving, Thomas R. and others. Air Force Enlisted Personnel Retention-Accession Model. Air Force Human Resources Laboratory (AFSC), Brooks AFB TX. June 1980. (AD-A085658)

Saving, Thomas R. and others. Retention of Air Force Enlisted Personnel: An Empirical Examination. Air Force Human Resources Laboratory (AFSC), Brooks AFB TX. July 1985. (AD-A158091)

Simpson, James R. A Methodology for Forecasting Voluntary Retention Rates for Air Force Pilots. MS thesis, AFIT/GOR-88M, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, December 1988.

Whalen, William P. An Analysis of Factors Affecting the Retention of Medical Officers in the United States Navy. MS thesis, Naval Postgraduate School, Monterrey CA, December 1986. (AD-A178588)

Woolard, David P. A Comparison of Variable Selection Criteria for Multiple Regression: A Second Simulation Study. MS thesis, AFIT/GOR-93M, School of Engineering, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, March 1993. (AD-A262512)

Vita

Captain Mark A. Basalla was born 24 September 1969 in Cleveland, Ohio. He graduated from Padua Franciscan High School in 1987. He attended Kent State University in Kent, Ohio with a four year Air Force Reserve Officer Training Corp (ROTC) scholarship. He graduated in 1991 with a B.S. in Applied Mathematics. After being commissioned, Mark was brought on active duty and assigned as a B-2 Advanced Technology Bomber Scientific Analyst at the 31st Test and Evaluation Squadron, Edwards AFB, CA. He was reassigned to the Air Force Institute of Technology in July 1994.

He married the former Kristen Elizabeth Engstli of Fairview Park, Ohio in 1992. They have one son, Brian - 2 years, and are expecting there second child in August.

Permanent Address: 1300 Mayview Ave.
                     Cleveland, Ohio 44109-3612